Parallel Text Retrieval and Wide Area Information Servers

Franklin Davis

Project Manager

Advanced Information Systems

Overview

- WAIS — Wide Area Information Servers
  - Project goals
  - WAIS system architecture
- Connection Machine Document Retrieval System
  - Relevance Feedback
  - Parallel text retrieval algorithms
  - Performance
- Future Systems

What is WAIS?

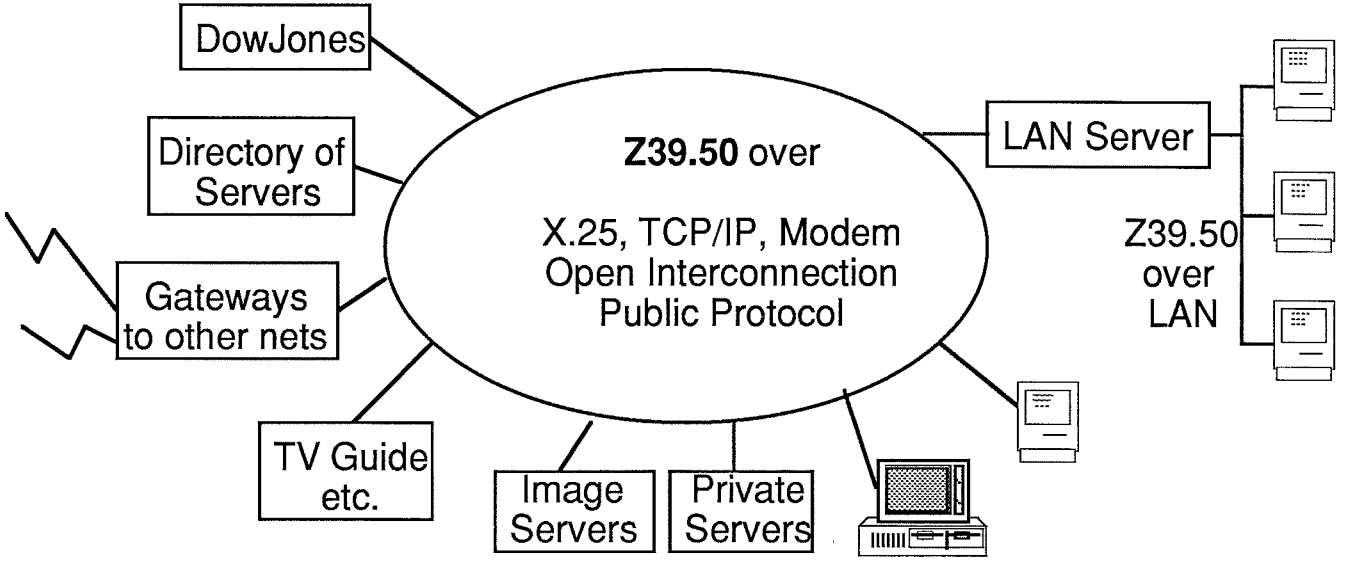
- WAIS Clients provide simple point-and-click interface
- Standard protocol connects clients to data servers
- WAIS Servers range from small to huge
- Improvements in servers and clients are independent

Levels of Information

- Personal files
  - Workgroup file server
    - Division database
      - Corporate/Organization database
        - Public databases

Goal: Access *all* levels from one interface

Wide Area Information Server Architecture

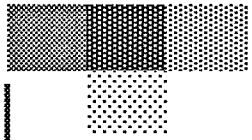


Users Needs:
- Automatically Selecting Servers
- Answering Questions
- Organizing Responses

Architecture Issues:
- Scalability
- Security
- Business model for servers
- Reliable Access

WAIS Clients

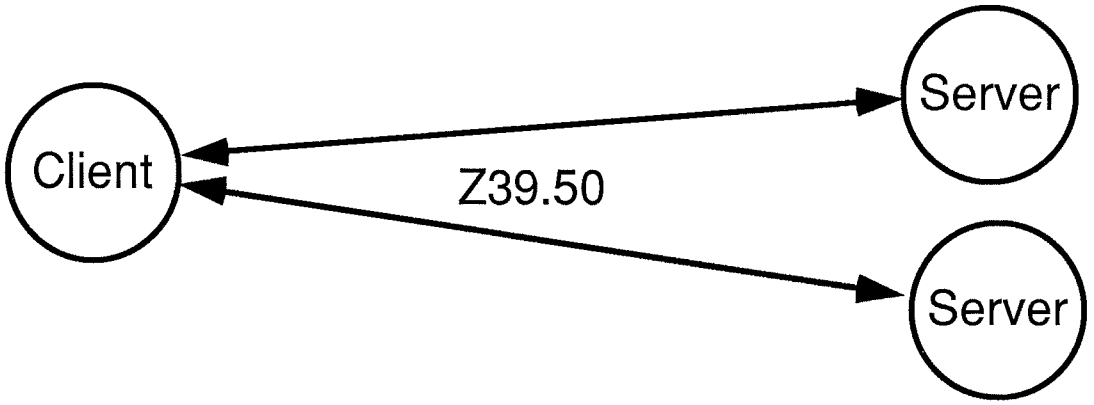
- Provide easy access to multiple sources
- Busy 24 hours a day finding information
- Automaticall learn user's preferences
- Scours the world (within a budget) to find new sources
- Current implementations on PC, Macintosh, X Windows, NeXT, dumb terminal



WAIS Protocol

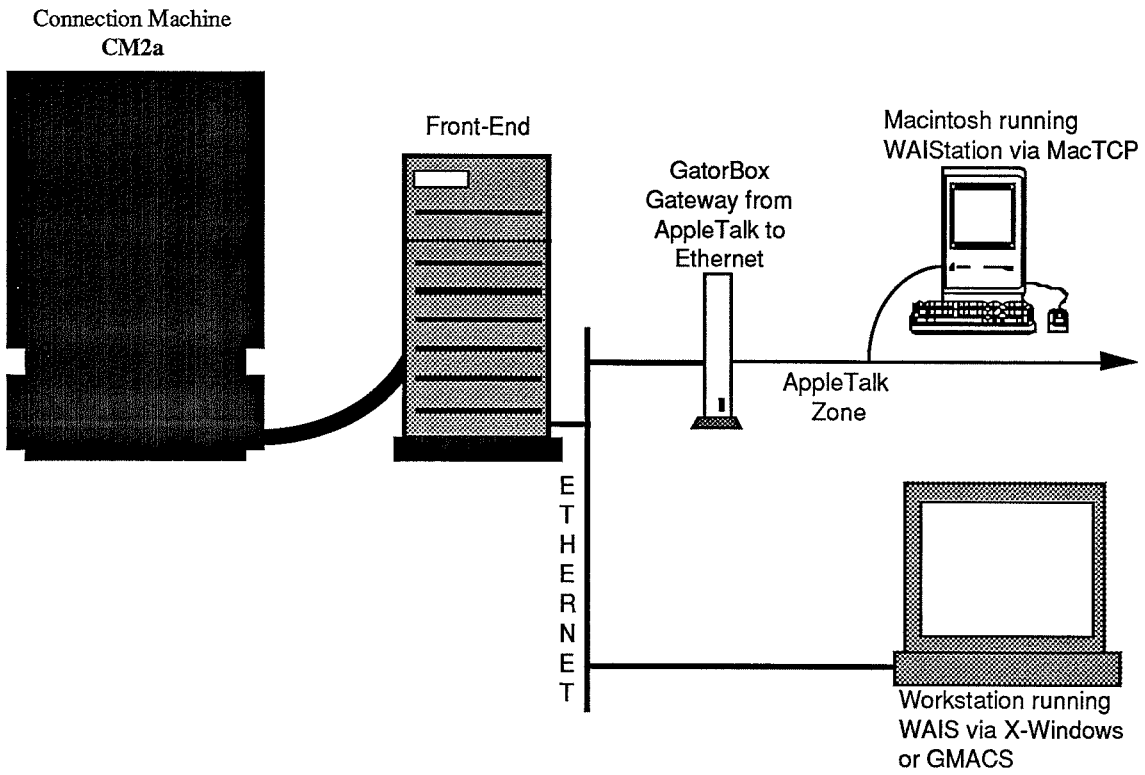
- Based on NISO Z39.50 international standard
- Flexible — separates clients from servers
- Search: (words, doc_ids, databases) returns list of: (headline, score, doc_id, types)
- Retrieval: (doc_id, type, start, end) returns: data of specified type

The WAIS Protocol *is* WAIS



- Supports any search syntax
- Supports sophisticated clients — puts intelligence in the user's hands
- Clients can run on any platform
- Multiple servers in a single search
- Retrieve any kind of data: text, graphics, video,…

WAIS Hardware Components



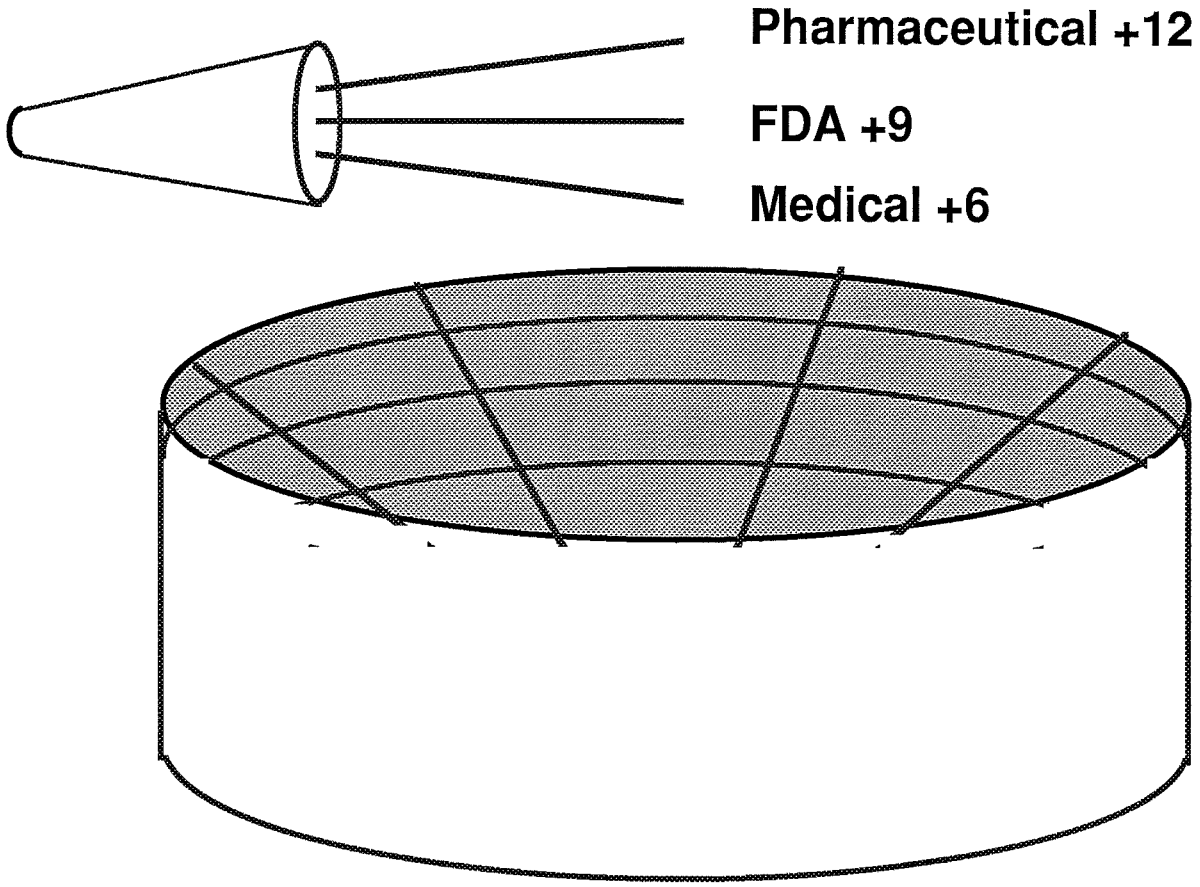
Connection Machine Server

- Interactive full-text retrieval with large queries
- 1-25 GBytes current CM-2 product
- Terabytes on future CM-5 system
- Supports thousands of users
- Automatic Indexing
- Uses words and phrases in question to find appropriate documents

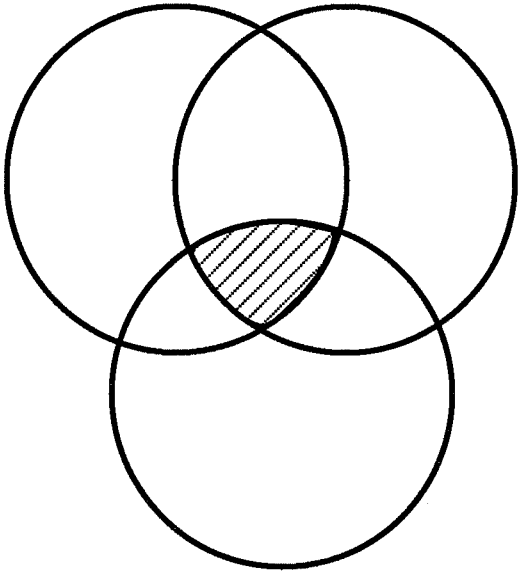
Why a Connection Machine?

- Bigger databases
- Interactive full-text search on gigabytes to terabytes
- More robust search techniques, e.g. relevance feedback, weighted terms

Data Parallelism:
Searching all the documents at once

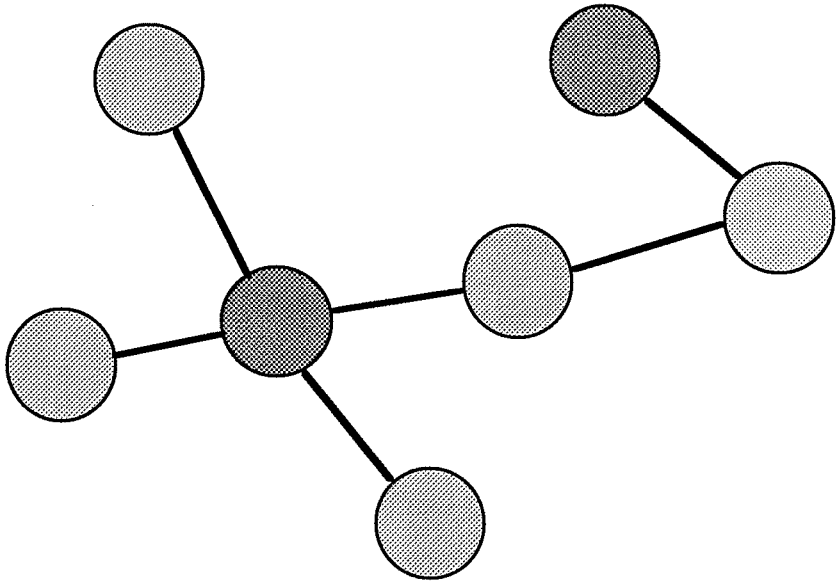


Boolean Search



Retrieve documents containing specific combinations of words

Conceptual Search



Explore a set of documents containing related concepts

Boolean Query

Hard to Use:
Complex Syntax

(Japanese OR Japan) AND
(building OR buildings OR (Real AND Estate) AND
(Manhattan OR (New AND York)

Poor Results:
The wrong information
No ranking of results

Have you been paying attention?...
Freer Finance: U.S. Regulators Move...
REAL ESTATE: California Initiatives...
First Boston Said To Agree on Sale Of...
Exxon, Rockefeller Group to Sell Site...
What's News--Business and Finance

Conceptual Search: Phase 1

Easy to Use:
No Syntax

Japanese buying real estate in mid-town manhattan

Options:

What do you want to follow up?

1. Time Acts to Cut Magazine Costs...
2. *First Boston Said To Agree on Sale...*
3. Have You Been Paying Attention?
4. *Exxon, Rockefeller Group to Sell Site...*
5. Hard Sell: Real Estate Developers...
6. What's News--Business and Finance...
7. Integrated Resources Buys Loft Building...

Conceptual Search: Phase 2

Relevance Feedback:

I like these; show me more

First Boston Said To Agree on Sale...
Exxon, Rockefeller Group to Sell Site...

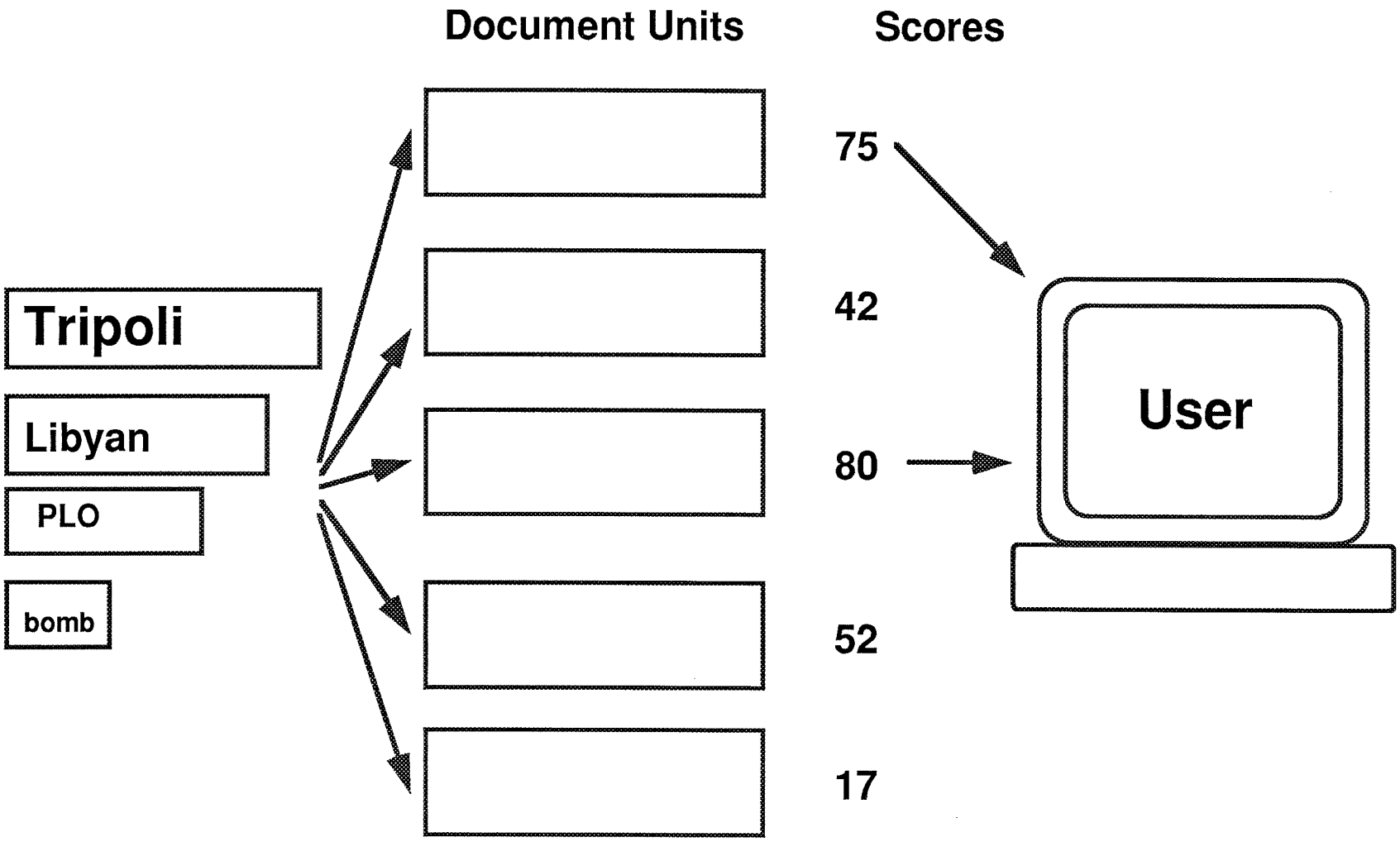
Improved results:

Articles on related topics are found

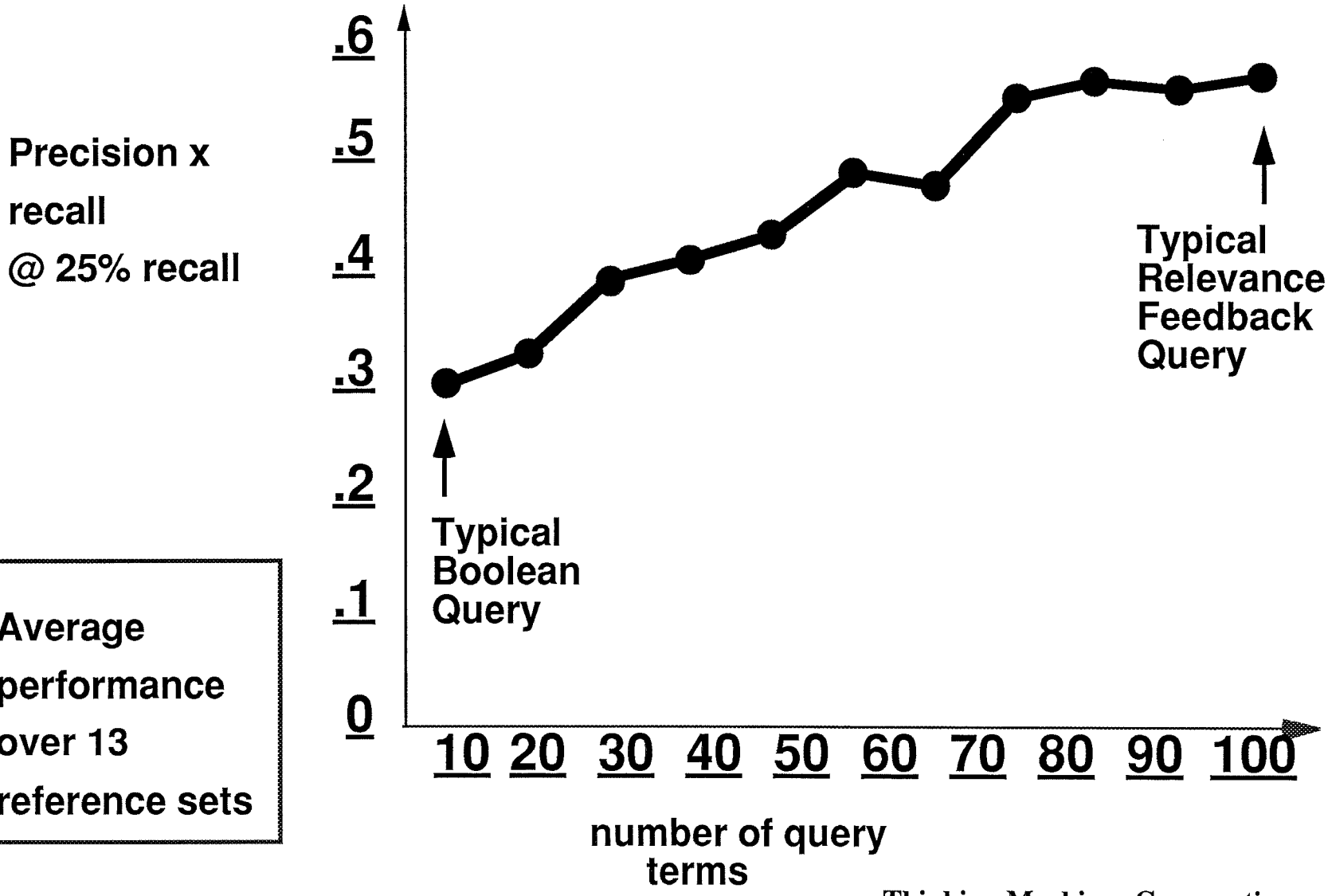
Results are ranked

1. Bids for Exxon Building in New York...
2. Time Acts to Cut Magazine Costs...
3. Hard Sell: Real Estate Developers...
4. Time Inc. Sells Its 45% Interest...
5. Citicorp Unit Moves to Foreclose on...
6. Litigious Landlords: Legal Maneuvers

Query Broadcast To Database on Connection Machine System



Results Improve with Query Size



Average performance over 13 reference sets

Document Retrieval Performance

- Current algorithm limits:
  - ~2 GB with 512 MB CM-2
  - ~8 GB with 2 GB CM-2
  - ~25 GB with 8 GB CM-2

- High recall
- High precision

} see Stanfill and Kahle, Communications of the ACM, December 1986

- << 1 sec. response

- Much larger DBs searchable with CM-5 and inverted index algorithms: 100s to 1000s of Gigabytes

DowQuest

- An advanced information retrieval service offered by Dow Jones News/Retrieval since January 1989
- Simple and powerful *search by example* model.
- Prime the system with a few words to find an article you like.
- Search again using good article: "Give me more articles like that."
- The full text database of over 400 publications is examined and compared with the reference article.
- 16 top scoring "best fit" articles retrieved almost instantly.
- Process is repeated until you find just the information you want.